Package 'scRNAseq'

October 24, 2025

Title Collection of Public Single-Cell RNA-Seq Datasets

Version 2.23.1

Date 2025-10-10

Description Gene-level counts for a collection of public scRNA-seq datasets, provided as SingleCellExperiment objects with cell- and gene-level metadata.

License CC0

NeedsCompilation no

Depends SingleCellExperiment

Imports utils, methods, Matrix, BiocGenerics, S4Vectors, SparseArray, DelayedArray, GenomicRanges, SummarizedExperiment, ExperimentHub (>= 2.3.4), AnnotationHub (>= 3.3.6), AnnotationDbi, ensembldb, GenomicFeatures, alabaster.base, alabaster.matrix, alabaster.sce, gypsum, jsonlite, DBI, RSQLite

Suggests BiocStyle, knitr, rmarkdown, testthat, jsonvalidate, BiocManager

VignetteBuilder knitr

Encoding UTF-8

biocViews ExperimentHub, ExperimentData, ExpressionData, SequencingData, RNASeqData, SingleCellData

BuildResaveData no

RoxygenNote 7.3.2

git_url https://git.bioconductor.org/packages/scRNAseq

git_branch devel

git_last_commit c47ea39

git_last_commit_date 2025-10-10

Repository Bioconductor 3.23

Date/Publication 2025-10-24

Author Davide Risso [aut, cph],
Michael Cole [aut],
Aaron Lun [ctb, cre],

Alan O'Callaghan [ctb],
Jens Preussner [ctb],
Charlotte Soneson [ctb],
Stephany Orjuela [ctb],
Daniel Bunis [ctb],
Milan Malfait [ctb]

Maintainer Aaron Lun <infinite.monkeys.with.keyboards@gmail.com>

Contents

`scRNAseq-package` *Collection of Public Single-Cell RNA-Seq Datasets*

Description

Gene-level counts for a collection of public scRNA-seq datasets, provided as SingleCellExperiment objects with cell- and gene-level metadata.

Details

This package contains a collection of three publicly available single-cell RNA-seq datasets.

The dataset `fluidigm` contains 65 cells from Pollen et al. (2014), each sequenced at high and low coverage.

The dataset `th2` contains 96 T helper cells from Mahata et al. (2014).

The dataset `allen` contains 379 cells from the mouse visual cortex. This is a subset of the data published in Tasic et al. (2016).

See the package vignette for details on the pre-processing of the data.

Author(s)

Davide Risso [aut, cph], Michael Cole [aut], Aaron Lun [ctb, cre], Alan O'Callaghan [ctb], Jens Preussner [ctb], Charlotte Soneson [ctb], Stephany Orjuela [ctb], Daniel Bunis [ctb], Milan Malfait [ctb]

Maintainer: Aaron Lun <infinite.monkeys.with.keyboards@gmail.com>

References

Pollen, Nowakowski, Shuga, Wang, Leyrat, Lui, Li, Szpankowski, Fowler, Chen, Ramalingam, Sun, Thu, Norris, Lebofsky, Toppani, Kemp II, Wong, Clerkson, Jones, Wu, Knutsson, Alvarado, Wang, Weaver, May, Jones, Unger, Kriegstein, West. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nature Biotechnology, 32, 1053-1058 (2014).

Mahata, Zhang, Kolodziejczyk, Proserpio, Haim-Vilmovsky, Taylor, Hebenstreit, Dingler, Moignard, Gottgens, Arlt, McKenzie, Teichmann. Single-Cell RNA Sequencing Reveals T Helper Cells Synthesizing Steroids De Novo to Contribute to Immune Homeostasis. Cell Reports, 7(4): 1130–1142 (2014).

Tasic, Menon, Nguyen, Kim, Jarsky, Yao, Levi, Gray, Sorensen, Dolbeare, Bertagnolli, Goldy, Shapovalova, Parry, Lee, Smith, Bernard, Madisen, Sunkin, Hawrylycz, Koch, Zeng. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. Nature Neuroscience, 19, 335–346 (2016).

`AztekinTailData` *Obtain the Aztekin tail data*

Description

Obtain the Xenopus tail single-cell RNA-seq data from Aztekin et al. (2019).

Usage

```
AztekinTailData(legacy = FALSE)
```

Arguments

`legacy`	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Column metadata is provided in the same form as supplied in E-MTAB-7761. This contains information such as the treatment condition, batch, putative cell type, putative cell cycle phase.

The UMAP results are available as the `"UMAP"` entry in the `reducedDims`.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/aztekin-tail`.

Value

A SingleCellExperiment object with a single matrix of UMI counts.

Author(s)

Aaron Lun

References

Aztekin C et al. (2019). Identification of a regeneration-organizing cell in the Xenopus tail. *Science* 364(6441), 653-658

Examples

```
sce <- AztekinTailData()
```

`BacherTCellData` *Obtain the Bacher T cell data*

Description

Obtain the human COVID T cell single-cell RNA-seq dataset from Bacher et al. (2020).

Usage

```
BacherTCellData(
  filtered = TRUE,
  ensembl = FALSE,
  location = TRUE,
  legacy = FALSE
)
```

Arguments

`filtered`	Logical scalar indicating whether to filter out cells that were not used by the authors.
`ensembl`	Logical scalar indicating whether the output row names should contain Ensembl identifiers.
`location`	Logical scalar indicating whether genomic coordinates should be returned.
`legacy`	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Column metadata is scraped from GEO, using both the author-supplied TSV of per-cell annotations and the sample-level metadata. This contains information such as the diagnosis, severity, WHO class, clustering and clonotype.

If `filtered=TRUE`, only the cells used by the authors in their final analysis are returned. Otherwise, an additional `filtered` field will be present in the `colData`, indicating whether the cell was retained by the authors.

If `ensembl=TRUE`, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

If `location=TRUE`, the coordinates of the Ensembl gene models are stored in the `rowRanges` of the output.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/bacher-tcell`.

Value

A SingleCellExperiment object with a single matrix of UMI counts.

Author(s)

Aaron Lun

References

Bacher P et al. (2020). Low avidity T cell responses to SARS-CoV-2 in unexposed individuals and severe COVID-19 *Immunity* 53, 1258-1271

Examples

```
if (.Machine$sizeof.pointer > 4) { # too large for 32-bit machines!
    sce <- BacherTCellData()
}
```

`BachMammaryData` *Obtain the Bach mammary data*

Description

Obtain the mouse mammary gland single-cell RNA-seq data from Bach et al. (2017).

Usage

```
BachMammaryData(
  samples = c("NP_1", "NP_2", "G_1", "G_2", "L_1", "L_2", "PI_1", "PI_2"),
  location = TRUE,
  legacy = FALSE
)
```

Arguments

samples	A character vector with at least one element, specifying which samples(s) to retrieve.
location	Logical scalar indicating whether genomic coordinates should be returned.
legacy	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Column metadata is extracted from the sample annotation in GSE106273, and refers to the developmental stage of the mammary gland.

If multiple samples are specified in `samples`, the count matrices will be `cbind`ed together. Cells originating from different samples are identifiable by the `"Sample"` field in the column metadata.

If `location=TRUE`, the coordinates of the Ensembl gene models are stored in the `rowRanges` of the output.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/bach-mammary`.

Value

A SingleCellExperiment object with a single matrix of UMI counts.

Author(s)

Aaron Lun

References

Bach K et al. (2017). Differentiation dynamics of mammary epithelial cells revealed by single-cell RNA sequencing. *Nat Commun.* 8(1), 2128

Examples

```
sce <- BachMammaryData(samples="NP_1")
```

`BaronPancreasData` *Obtain the Baron pancreas data*

Description

Obtain the human/mouse pancreas single-cell RNA-seq data from Baron et al. (2017).

Usage

```
BaronPancreasData(
  which = c("human", "mouse"),
  ensembl = FALSE,
  location = TRUE,
  legacy = FALSE
)
```

Arguments

`which`	String specifying the species to get data for.
`ensembl`	Logical scalar indicating whether the output row names should contain Ensembl identifiers.
`location`	Logical scalar indicating whether genomic coordinates should be returned.
`legacy`	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Column metadata is provided in the same form as supplied in GSE84133. This contains information such as the cell type labels and donor ID (for humans) or strain (for mouse).

If `ensembl=TRUE`, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

If `location=TRUE`, the coordinates of the Ensembl gene models are stored in the `rowRanges` of the output. Note that this is only performed if `ensembl=TRUE`.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/baron-pancreas`.

Value

A SingleCellExperiment object with a single matrix of read counts.

Author(s)

Aaron Lun

References

Baron M et al. (2017). Single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Syst.* 3(4), 346-360.

Examples

```
sce.human <- BaronPancreasData()

sce.mouse <- BaronPancreasData("mouse")
```

`BhaduriOrganoidData` *Obtain the Bhaduri cortical organoid data*

Description

Obtain the human cortical organoid single-cell RNA-seq dataset from Bhaduri et al. (2020).

Usage

```
BhaduriOrganoidData(ensembl = FALSE, location = TRUE, legacy = FALSE)
```

Arguments

`ensembl`	Logical scalar indicating whether the output row names should contain Ensembl identifiers.
`location`	Logical scalar indicating whether genomic coordinates should be returned.
`legacy`	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Column data contains sample-level information. In theory, there is also cell-level metadata for this dataset but it could not be unambiguously mapped to the column names.

If `ensembl=TRUE`, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

If `location=TRUE`, the coordinates of the Ensembl gene models are stored in the `rowRanges` of the output. Note that this is only performed if `ensembl=TRUE`.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/bhaduri-organoid`.

Value

A SingleCellExperiment object with a single matrix of normalized expression values.

Author(s)

Aaron Lun

References

Bhaduri A et al. (2020). Cell stress in cortical organoids impairs molecular subtype specification. *Nature* 578(7793), 142-148.

Examples

```
if (.Machine$sizeof.pointer > 4) { # too large for 32-bit machines!
    sce <- BhaduriOrganoidData()
}
```

`BuettnerESCData` *Obtain the Buettner ESC data*

Description

Obtain the mouse embryonic stem cell single-cell RNA-seq data from Buettner et al. (2015).

Usage

```
BuettnerESCData(remove.htseq = TRUE, location = TRUE, legacy = FALSE)
```

Arguments

`remove.htseq`	Logical scalar indicating whether HT-seq alignment statistics should be removed.
`location`	Logical scalar indicating whether genomic coordinates should be returned.
`legacy`	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Rows corresponding to HT-seq's alignment statistics are removed by default. These can be retained by setting `remove.htseq=FALSE`.

Column metadata contains the experimentally determined cell cycle phase for each cell.

Counts for ERCC spike-ins are stored in the `"ERCC"` entry in the `altExps`.

If `location=TRUE`, the coordinates of the Ensembl gene models are stored in the `rowRanges` of the output.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/buettner-esc`.

Value

A SingleCellExperiment object with a single matrix of read counts.

Author(s)

Aaron Lun

References

Buettner F et al. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* 33(2), 155-160.

Examples

```
sce <- BuettnerESCData()
```

BunisHSPCData *Obtain the Bunis haematopoietic stem and progenitor cell data*

Description

Obtain the human fetal, newborn, and adult haematopoietic stem and progenitor cell single-cell RNA-seq dataset from Bunis et al. (2021).

Usage

```
BunisHSPCData(filtered = TRUE, legacy = FALSE)
```

Arguments

`filtered`	Logical scalar or "cells" indicating whether to: • `TRUE`: filter out cells that were not used by the authors. • `"cells"`: filter out empty droplets as filtered out by cell ranger. • `FALSE`: no filtering
`legacy`	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Column metadata is recreated from GEO using the author-supplied TSV of per-cell annotations, or retrieved from a processed version of the data shared by authors via figshare. This contains information such as the tissue & sample of origin, age group, likely cell type, and Developmental Stage Scoring. Within DevStageScoring element of the column metadata are the applied results ('<cell_type>_scores') of random forest regression trained on the fetal (score = 0) and adult (score = 1) cells of individual cell types indicated by ('<cell_type>_inTraining').

If `filtered=TRUE`, only the cells used by the authors in their final analysis are returned. Otherwise, an additional `retained` field will be present in the `colData`, indicating whether the cell was retained by the authors.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/bunis-hspc`.

Value

A SingleCellExperiment object with a single matrix of UMI counts.

Author(s)

Daniel Bunis

References

Bunis DG et al. (2021). Single-Cell Mapping of Progressive Fetal-to-Adult Transition in Human Naive T Cells *Cell Rep.* 34(1): 108573

Examples

```
sce <- BunisHSPCData()
```

`CampbellBrainData` *Obtain the Campbell brain data*

Description

Obtain the mouse brain single-cell RNA-seq data from Campbell et al. (2017).

Usage

```
CampbellBrainData(ensembl = FALSE, location = TRUE, legacy = FALSE)
```

Arguments

`ensembl`	Logical scalar indicating whether the row names of the returned object should contain Ensembl identifiers.
`location`	Logical scalar indicating whether genomic coordinates should be returned.
`legacy`	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Column metadata is provided in the same form as supplied in GSE93374. This contains information such as the diet of the mice, sex and proposed cell type for each cell.

If `ensembl=TRUE`, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

If `location=TRUE`, the coordinates of the Ensembl gene models are stored in the `rowRanges` of the output. Note that this is only performed if `ensembl=TRUE`.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/campbell-brain`.

Value

A SingleCellExperiment object with a single matrix of UMI counts.

Author(s)

Aaron Lun

References

Campbell R et al. (2017). A molecular census of arcuate hypothalamus and median eminence cell types. *Nat. Neurosci.* 20, 484-496.

Examples

```
sce <- CampbellBrainData()
```

`ChenBrainData` *Obtain the Chen brain data*

Description

Obtain the mouse brain single-cell RNA-seq data from Chen et al. (2017).

Usage

```
ChenBrainData(ensembl = FALSE, location = TRUE, legacy = FALSE)
```

Arguments

`ensembl`	Logical scalar indicating whether the output row names should contain Ensembl identifiers.
`location`	Logical scalar indicating whether genomic coordinates should be returned.
`legacy`	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Column metadata is provided in the same form as supplied in GSE87544. This contains the putative cell type assigned by the original authors.

If `ensembl=TRUE`, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

If `location=TRUE`, the coordinates of the Ensembl gene models are stored in the `rowRanges` of the output. Note that this is only performed if `ensembl=TRUE`.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/chen-brain`.

Value

A SingleCellExperiment object with a single matrix of UMI counts.

Author(s)

Aaron Lun

References

Chen R et al. (2017). Single-Cell RNA-Seq reveals hypothalamic cell diversity. *Cell Rep.* 18, 3227-3241.

Examples

```
sce <- ChenBrainData()
```

`countErccMolecules` *Obtain ERCC molecule counts from concentrations*

Description

Compute the number of molecules for each transcript in the ERCC spike-in mixture, based on their published concentration as well as the volume of the diluted mixture added to each cell.

Usage

```
countErccMolecules(volume, dilution, mix = c("1", "2"), ...)
```

Arguments

`volume`	Numeric scalar specifying the added volume (in microliters) of ERCC spike-in mixture.
`dilution`	Numeric scalar specifying the dilution factor used for the added volume of the spike-in mixture.
`mix`	String specifying whether to compute the number of molecules for mix 1 or 2.
`...`	Further arguments to pass to `fetchDataset`.

Value

A DataFrame object with one row per ERCC spike-in transcript. This contains the estimated concentration and molecule count for each transcript.

Author(s)

Aaron Lun, based on code from Alan O'Callaghan

Examples

```
countErccMolecules(volume = 9, dilution = 300000)
```

`DarmanisBrainData` *Obtain the Darmanis brain data*

Description

Obtain the human brain single-cell RNA-seq dataset from Darmanis et al. (2015).

Usage

```
DarmanisBrainData(
  ensembl = FALSE,
  location = TRUE,
  remove.htseq = TRUE,
  legacy = FALSE
)
```

Arguments

`ensembl`	Logical scalar indicating whether the output row names should contain Ensembl identifiers.
`location`	Logical scalar indicating whether genomic coordinates should be returned.
`remove.htseq`	Logical scalar indicating whether HT-seq alignment statistics should be removed.
`legacy`	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Column metadata is scraped from GEO and includes patient information, tissue of origin and likely cell type.

If `ensembl=TRUE`, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

If `location=TRUE`, the coordinates of the Ensembl gene models are stored in the `rowRanges` of the output. This is only performed when `ensembl=TRUE`.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/darmanis-brain`.

Value

A SingleCellExperiment object with a single matrix of UMI counts.

Author(s)

Aaron Lun

References

Darmanis S et al. (2015). A survey of human brain transcriptome diversity at the single cell level. *Proc Natl Acad Sci USA* 112, 7285-90.

Examples

```
sce <- DarmanisBrainData()
```

ERCCSpikeInConcentrations

Obtain ERCC concentrations

Description

Obtain ERCC spike-in concentrations from the Thermo Fisher Scientific website.

Usage

```
ERCCSpikeInConcentrations(
  volume = NULL,
  dilution = NULL,
  mix = c("1", "2"),
  legacy = FALSE
)
```

Arguments

`volume`	Numeric scalar specifying the added volume (in nanoliters) of ERCC spike-in mixture. Only used if `dilution` is specified.
`dilution`	Numeric scalar specifying the dilution factor used for the added volume of the spike-in mixture. Only used if `volume` is specified.
`mix`	String specifying whether to compute the number of molecules for mix 1 or 2. Only used if both `dilution` and `volume` are specified.
`legacy`	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

If `volume` and `dilution` are specified, an additional column is added to the output specifying the number of molecules of spike-in transcipt for the specified mix.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/ercc-concentrations`.

Value

A DataFrame object with one row per ERCC spike-in transcript. This contains information such as the spike-in concentration in each mix.

Author(s)

Alan O'Callaghan

Examples

```
df <- ERCCSpikeInConcentrations()
```

`ErnstSpermatogenesisData` *Obtain the Ernst spermatogenesis data*

Description

Obtain the mouse spermatogenesis single-cell RNA-seq dataset from Ernst et al. (2019).

Usage

```
ErnstSpermatogenesisData(
  method = c("emptyDrops", "Cellranger"),
  location = TRUE,
  legacy = FALSE
)
```

Arguments

`method`	String indicating which cell caller to obtain results for.
`location`	Logical scalar indicating whether genomic coordinates should be returned.
`legacy`	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

This study contains two analyses done with datasets from different cell calling algorithms. One uses Cellranger version 2 while the other uses `emptyDrops` from **DropletUtils**.

Column metadata includes sample information, per-cell QC metrics and cell type labels. In particular, the sample label specifies the developmental stage of the mouse.

Note that `method="Cellranger"` contains additional data for Tc1 mice. These mice have an additional human chromosome 21 inserted alongside the usual mouse chromosomes.

If `location=TRUE`, the coordinates of the Ensembl gene models are stored in the `rowRanges` of the output.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/ernst-spermatogenesis`.

Value

A SingleCellExperiment object with a single matrix of UMI counts.

Author(s)

Aaron Lun

References

Ernst C et al. (2019). Staged developmental mapping and X chromosome transcriptional dynamics during mouse spermatogenesis. *Nat Commun* 10, 1251

Examples

```
if (.Machine$sizeof.pointer > 4) { # too large for 32-bit machines!
    sce <- ErnstSpermatogenesisData()
}
```

---

fetchDataset *Fetch a dataset from the gypsum backend*

Description

Fetch a dataset (or its metadata) from the gypsum backend.

Usage

```
fetchDataset(
  name,
  version,
  path = NA,
  package = "scRNAseq",
  cache = cacheDirectory(),
  overwrite = FALSE,
  realize.assays = FALSE,
  realize.reduced.dims = TRUE,
  ...
)

fetchMetadata(
  name,
  version,
  path = NA,
  package = "scRNAseq",
  cache = cacheDirectory(),
  overwrite = FALSE
)
```

Arguments

name	String containing the name of the dataset.
version	String containing the version of the dataset.
path	String containing the path to a subdataset, if name consists of multiple subdatasets. Defaults to NA if no subdatasets are present.
package	String containing the name of the package.
cache, overwrite	Arguments to pass to saveVersion or saveFile.
realize.assays, realize.reduced.dims	Logical scalars indicating whether to realize assays and reduced dimensions into memory. Dense and sparse ReloadedArray objects are converted into ordinary arrays and dgCMatrix objects, respectively.
...	Further arguments to pass to readObject.

Value

fetchDataset returns the dataset as a SummarizedExperiment or one of its subclasses.

fetchMetadata returns a named list of metadata for the specified dataset.

Author(s)

Aaron Lun

See Also

https://github.com/ArtifactDB/bioconductor-metadata-index, on the expected schema for the metadata.

saveDataset and uploadDirectory, to save and upload a dataset.

surveyDatasets and listVersions, to get possible values for name and version.

Examples

```
fetchDataset("zeisel-brain-2015", "2023-12-14")
fetchMetadata("zeisel-brain-2015", "2023-12-14")
```

FletcherOlfactoryData *Obtain the Fletcher Olfactory data*

Description

Obtain the mouse olfactory epithelial HBC stem cell differentiation dataset from Fletcher et al. (2017).

Usage

```
FletcherOlfactoryData(
  filtered = TRUE,
  ensembl = FALSE,
  location = TRUE,
  legacy = FALSE
)
```

Arguments

`filtered`	Logical scalar indicating whether to filter out cells that were not used by the authors.
`ensembl`	Logical scalar indicating whether the output row names should contain Ensembl identifiers.
`location`	Logical scalar indicating whether genomic coordinates should be returned.
`legacy`	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Column metadata is scraped from GEO, using both the author-supplied "phenoData" per-cell annotations and the author-supplied "protocolData" per-cell annotations. The former includes information about the animals and the instruments, while the latter contains QC statistics.

We also included the clustering results from the authors' analysis.

If `filtered=TRUE`, only the cells used by the authors in their cluster analysis are returned. Otherwise, the cells not used by the authors will have `NA` in the clustering columns of the `colData`.

If `ensembl=TRUE`, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

If `location=TRUE`, the coordinates of the Ensembl gene models are stored in the `rowRanges` of the output.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/fletcher-olfactory`.

Value

A SingleCellExperiment object with a single matrix of read counts.

Author(s)

Davide Risso

References

Fletcher R et al. (2017). Deconstructing olfactory stem cell trajectories at single-cell resolution. *Cell Stem Cell* 20, 817-30.

Examples

```
sce <- FletcherOlfactoryData()
```

`GiladiHSCData` *Obtain the Giladi HSC data*

Description

Obtain the mouse haematopoietic stem cell single-cell RNA-seq and CRISPR-seq dataset from Giladi et al. (2018).

Usage

```
GiladiHSCData(
  mode = c("rna", "crispr"),
  filtered = TRUE,
  ensembl = FALSE,
  location = TRUE,
  legacy = FALSE
)
```

Arguments

`mode`	Character vector specifying which modalities should be returned.
`filtered`	Logical scalar indicating whether to filter out cells that were not used by the authors.
`ensembl`	Logical scalar indicating whether the output row names should contain Ensembl identifiers, when `mode` contains `"rna"`.
`location`	Logical scalar indicating whether genomic coordinates should be returned, when `mode` contains `"rna"`.
`legacy`	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Column metadata is scraped from GEO using the author-supplied TSV of per-cell annotations. This contains information such as the batch of origin for each cell plus an array of FACS measurements per cell.

If `filtered=TRUE`, only the cells used by the authors in their final analysis are returned. Otherwise, an additional `filtered` field will be present in the `colData`, indicating whether the cell was retained by the authors.

If `ensembl=TRUE`, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated

IDs is retained. For row names with multiple semi-colon-delimited symbols, the last symbol is used for matching against the Ensembl annotation.

If `location=TRUE`, the coordinates of the Ensembl gene models are stored in the `rowRanges` of the output. This is only relevant when `ensembl=TRUE`.

If `mode` contains multiple modalities, the intersection of cells that are present in both modalities is returned. This is because not all cells have data across both modalities. If `mode` contains only one modality, all cells for that modality are returned.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/giladi-hsc`.

Value

A SingleCellExperiment object with a matrix of UMI counts for the scRNA-seq or CRISPR-seq data. Alternatively, an object with both count matrices, where the second modality is stored as an alternative Experiment.

Author(s)

Aaron Lun

References

Giladi A et al. (2018). Single-cell characterization of haematopoietic progenitors and their trajectories in homeostasis and perturbed haematopoiesis. *Nat Cell Biol* 20, 836-846

Examples

```
if (.Machine$sizeof.pointer > 4) { # too large for 32-bit machines!
    sce <- GiladiHSCData()
}
```

`GrunHSCData` *Obtain the Grun HSC data*

Description

Obtain the mouse haematopoietic stem cell single-cell RNA-seq data from Grun et al. (2016).

Usage

```
GrunHSCData(ensembl = FALSE, location = TRUE, legacy = FALSE)
```

Arguments

`ensembl`	Logical scalar indicating whether the output row names should contain Ensembl identifiers.
`location`	Logical scalar indicating whether genomic coordinates should be returned.
`legacy`	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Row metadata contains the symbol and chromosomal location for each gene. Column metadata contains the extraction protocol used for each sample, as described in GSE76983.

If `ensembl=TRUE`, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

If `location=TRUE`, the coordinates of the Ensembl gene models are stored in the `rowRanges` of the output. Note that this is only performed if `ensembl=TRUE`.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/grun-hsc`.

Value

A SingleCellExperiment object with a single matrix of UMI counts.

Author(s)

Aaron Lun

References

Grun D et al. (2016). De novo prediction of stem cell identity using single-cell transcriptome data. *Cell Stem Cell* 19(2), 266-277.

Examples

```
sce <- GrunHSCData()
```

`GrunPancreasData` *Obtain the Grun pancreas data*

Description

Obtain the human pancreas single-cell RNA-seq data from Grun et al. (2016).

Usage

```
GrunPancreasData(ensembl = FALSE, location = TRUE, legacy = FALSE)
```

Arguments

`ensembl`	Logical scalar indicating whether the output row names should contain Ensembl identifiers.
`location`	Logical scalar indicating whether genomic coordinates should be returned.
`legacy`	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Row metadata contains fields for the symbol and chromosomal location of each gene, as derived from the row names.

Column metadata is derived from the column names of the count matrix with the sample annotations in GSE81076. This includes the donor identity for each cell and the type of sample.

The `"ERCC"` entry in the `altExps` contains count data for the ERCC spike-in transcripts.

If `ensembl=TRUE`, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

If `location=TRUE`, the coordinates of the Ensembl gene models are stored in the `rowRanges` of the output. Note that this is only performed if `ensembl=TRUE`.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/grun-pancreas`.

Value

A SingleCellExperiment object with a single matrix of UMI counts.

Author(s)

Aaron Lun, using additional metadata obtained by Vladimir Kiselev.

References

Grun D et al. (2016). De novo prediction of stem cell identity using single-cell transcriptome data. *Cell Stem Cell* 19(2), 266-277.

Examples

```
sce <- GrunPancreasData()
```

`HeOrganAtlasData` *Obtain the He organ atlas data*

Description

Obtain the human cortex single-nuclei RNA-seq data from Hu et al. (2017).

Usage

```
HeOrganAtlasData(
  tissue = c("Bladder", "Blood", "Common.bile.duct", "Esophagus", "Heart", "Liver",
    "Lymph.node", "Marrow", "Muscle", "Rectum", "Skin", "Small.intestine", "Spleen",
      "Stomach", "Trachea"),
  ensembl = FALSE,
  location = TRUE,
  legacy = FALSE
)
```

Arguments

tissue	Character vector specifying the tissues to return.
ensembl	Logical scalar indicating whether the output row names should contain Ensembl identifiers.
location	Logical scalar indicating whether genomic coordinates should be returned.
legacy	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Column data contains the tissue of origin, a variety of per-cell QC metrics well as some cell type annotations. The reclustered annotations required some assembly:

- `reclustered.broad` was generated based on whether the barcode was present in each `*_meta.data.txt` file at https://github.com/bei-lab/scRNA-AHCA.
- For each barcode that was present in one of those files, `reclustered.fine` was generated based on the label in the `annotation` field inside that file.

If multiple tissues are requested, counts are only reported for the intersection of genes across all tissues. This is because the gene annotation in the original count matrices differs across tissues.

If `ensembl=TRUE`, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

If `location=TRUE`, the coordinates of the Ensembl gene models are stored in the `rowRanges` of the output. Note that this is only performed if `ensembl=TRUE`.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/he-organ-atlas`.

Value

A SingleCellExperiment object with a single matrix of UMI counts.

Author(s)

Aaron Lun

References

He S et al. (2020). Single-cell transcriptome profiling of an adult human cell atlas of 15 major organs. *Genome Biol* 21, 1:294.

Examples

```
if (.Machine$sizeof.pointer > 4) { # too large for 32-bit machines!
    sce <- HeOrganAtlasData()
}
```

`HermannSpermatogenesisData` *Obtain the Hermann spermatogenesis data*

Description

Obtain the mouse spermatogenesis single-cell RNA-seq data from Hermann et al. (2018).

Usage

```
HermannSpermatogenesisData(strip = FALSE, location = TRUE, legacy = FALSE)
```

Arguments

`strip`	Logical scalar indicating whether to strip the `.X` notation from the row names.
`location`	Logical scalar indicating whether genomic coordinates should be returned. Only used if `strip=TRUE`.
`legacy`	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Column metadata contains cell types provided by the data generators at https://data.mendeley.com/datasets/kxd5f8vpt4/1#file-fe79c10b-c42e-472e-9c7e-9a9873d9b3d8.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/hermann-spermatogenesis`.

Value

A SingleCellExperiment object with two matrices, containing spliced and unspliced counts, respectively.

Author(s)

Charlotte Soneson

References

Hermann B.P. et al. (2018). The Mammalian Spermatogenesis Single-Cell Transcriptome, from Spermatogonial Stem Cells to Spermatids. *Cell Rep.* 25(6), 1650-1667.e8.

Examples

```
sce <- HermannSpermatogenesisData()
```

`HuCortexData` *Obtain the Hu cortex data*

Description

Obtain the mouse cortex single-nuclei RNA-seq data from Hu et al. (2017).

Usage

```
HuCortexData(
  mode = c("ctx", "3T3"),
  samples = NULL,
  ensembl = FALSE,
  location = TRUE,
  legacy = FALSE
)
```

Arguments

`mode`	Character vector indicating whether to return data for the 3T3 cells or the mouse cortex.
`samples`	Character vector indicating whether to return data for specific samples, see Details. If specified, this overrides `mode`.
`ensembl`	Logical scalar indicating whether the output row names should contain Ensembl identifiers.
`location`	Logical scalar indicating whether genomic coordinates should be returned.
`legacy`	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Column metadata includes the mode and sample corresponding to each cell/nuclei. Available samples are:

- `"cell-3T3"` and `"nuclei-3T3"`, generated from the 3T3 cell line.
- `"nuclei-ctx-X"`, nuclei generated from the cortex of animal number X (from 1 to 13).

- "nuclei-ctx-salineX" or "nuclei-ctx-PTZX", nuclei generated from the cortex of saline- or PTZ-treated mice. X represents the replicate number and can be 1 or 2.

If multiple modes are requested, counts are only reported for the intersection of genes across all modes. This is because the gene annotation in the original count matrices differs across modes.

If ensembl=TRUE, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

If location=TRUE, the coordinates of the Ensembl gene models are stored in the rowRanges of the output. Note that this is only performed if ensembl=TRUE.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for scRNAseq/wu-kidney.

Value

A SingleCellExperiment object with a single matrix of read counts.

Author(s)

Aaron Lun

References

Hu P et al. (2017). Dissecting cell-type composition and activity-dependent transcriptional state in mammalian brains by massively parallel single-nucleus RNA-seq. *Mol. cell* 68, 1006-1015.

Examples

```
sce <- HuCortexData("3T3")
```

JessaBrainData *Obtain the Jessa brain data*

Description

Obtain the mouse brain single-cell RNA-seq dataset from Jessa et al. (2019).

Usage

```
JessaBrainData(filtered = TRUE, location = TRUE, legacy = FALSE)
```

Arguments

filtered	Logical scalar indicating whether to filter out cells that were not used by the authors.
location	Logical scalar indicating whether genomic coordinates should be returned.
legacy	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

If `filtered=TRUE`, only the cells used by the authors in their final analysis are returned. Otherwise, an additional `filtered` field will be present in the `colData`, indicating whether the cell was retained by the authors.

The column data contains sample of origin, some QC metrics and various cluster assignments for each cell. Cluster assignments starting with `Sample_*` are derived from per-sample analyses and cannot be compared sensibly across samples. Other clusterings (`Forebrain_*` and `Pons_*`) are derived from joint analyses across all samples involving the named tissue.

The `reducedDims` of the output contains various dimensionality reduction results. Coordinates for entries prefixed with `Sample_*` were generated from per-sample analyses and cannot be compared across samples. Coordinates for entries prefixed with `Forebrain_*` and `Pons_*` were generated from joint analyses from the corresponding tissue.

If `location=TRUE`, the coordinates of the Ensembl gene models are stored in the `rowRanges` of the output.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/jessa-brain`.

Value

A SingleCellExperiment object with a single matrix of UMI counts.

Author(s)

Aaron Lun

References

Jessa S et al. (2019). Stalled developmental programs at the root of pediatric brain tumors *Nat Genet* 51, 1702-1713

Examples

```
if (.Machine$sizeof.pointer > 4) { # too large for 32-bit machines!
    sce <- JessaBrainData()
}
```

`KolodziejczykESCData` *Obtain the Kolodziejcyzk ESC data*

Description

Obtain the mouse embryonic stem cell single-cell RNA-seq data from Kolodziejczyk et al. (2015).

Usage

```
KolodziejczykESCData(remove.htseq = TRUE, location = TRUE, legacy = FALSE)
```

Arguments

remove.htseq	Logical scalar indicating whether HT-seq alignment statistics should be removed.
location	Logical scalar indicating whether genomic coordinates should be returned.
legacy	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Column metadata is generated from the column names, and contains the culture conditions and the plate of origin for each cell.

Count data for ERCC spike-ins are stored in the "ERCC" entry in the altExps.

If location=TRUE, the coordinates of the Ensembl gene models are stored in the rowRanges of the output.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for scRNAseq/kolodziejczyk-esc.

Value

A SingleCellExperiment object with a single matrix of read counts.

Author(s)

Aaron Lun

References

Messmer T et al. (2019). Transcriptional heterogeneity in naive and primed human pluripotent stem cells at single-cell resolution. *Cell Rep* 26(4), 815-824.e4

Examples

```
sce <- KolodziejczykESCData()
```

KotliarovPBMCData *Obtain the Kotliarov CITE-seq data*

Description

Obtain the Kotliarov PBMC CITE-seq data from Kotliarov et al. (2020).

Usage

```
KotliarovPBMCData(
  mode = c("rna", "adt"),
  ensembl = FALSE,
  location = TRUE,
  legacy = FALSE
)
```

Arguments

mode	Character vector specifying whether to return either or both the RNA and ADT counts.
ensembl	Logical scalar indicating whether the output row names should contain Ensembl identifiers.
location	Logical scalar indicating whether genomic coordinates should be returned.
legacy	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

This dataset contains 20 samples from 2 experimental batches, where each batch contains 5 high and 5 low responders. The 10 samples per batch were mixed and distributed across the 6 lanes using a cell hashing approach.

The column metadata contains the following fields:

- `sample*`: identifiers for the sample of origin for each cell.
- `adjmfc.time`: type of responder for each sample.
- `tenx_lane`: 10X lane from which each cell was collected.
- `batch`: the batch of origin.
- `barcode_check`: barcode identifier.
- `hash_*` and `hto_*` columns: **HTOdemux** outputs.
- `DEMUXLET.*` columns: **demuxlet** outputs.
- `joint_classification_global`: **HTOdemux** and **demuxlet** joint classification.
- `nGene`: number of genes as defined from **Seurat**'s `CreateSeuratObject`.
- `nUMI`: number of UMIs as defined from **Seurat**'s `CreateSeuratObject`.
- `pctMT`: percent of mitochondrial reads as defined from **Seurat**'s `CreateSeuratObject`.

Note, no filtering has been performed based on the quality control metrics.

If `ensembl=TRUE`, the gene symbols in the RNA data are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

If `location=TRUE`, the coordinates of the Ensembl gene models are stored in the `rowRanges` for the RNA data. Note that this is only performed if `ensembl=TRUE`.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/kotliarov-pbmc`.

Value

A SingleCellExperiment object with a single matrix of UMI counts corresponding to the first mode, with an optional alternative Experiment if there is a second mode.

Author(s)

Stephany Orjuela, with modifications from Aaron Lun

References

Kotliarov et al. (2020). Broad immune activation underlies shared set point signatures for vaccine responsiveness in healthy individuals and disease activity in patients with lupus. *Nat. Med.* 26, 618–629

Examples

```
sce <- KotliarovPBMCData()
```

LaMannoBrainData *Obtain the La Manno brain data*

Description

Obtain the mouse/human brain scRNA-seq data from La Manno et al. (2016).

Usage

```
LaMannoBrainData(
  which = c("human-es", "human-embryo", "human-ips", "mouse-adult", "mouse-embryo"),
  ensembl = FALSE,
  location = TRUE,
  legacy = FALSE
)
```

Arguments

which	A string specifying which dataset should be obtained.
ensembl	Logical scalar indicating whether the output row names should contain Ensembl identifiers.
location	Logical scalar indicating whether genomic coordinates should be returned.
legacy	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Column metadata is provided in the same form as supplied in the supplementary tables in GSE71585. This contains information such as the time point and cell type.

The various settings of `which` will obtain different data sets.

- `"human-es"`, human embryonic stem cells.
- `"human-embryo"`, human embryo midbrain.
- `"human-ips"`, human induced pluripotent stem cells.
- `"mouse-adult"`, mouse adult dopaminergic neurons.
- `"mouse-embryo"`, mouse embryo midbrain.

Unfortunately, each of these datasets uses a different set of features. If multiple datasets are to be used simultaneously, users will have to decide how to merge them, e.g., by taking the intersection of common features across all datasets.

If `ensembl=TRUE`, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

If `location=TRUE`, the coordinates of the Ensembl gene models are stored in the `rowRanges` of the output. Note that this is only performed if `ensembl=TRUE`.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/lamanno-brain`.

Value

A SingleCellExperiment object with a single matrix of UMI counts.

Author(s)

Aaron Lun

References

La Manno A et al. (2016). Molecular diversity of midbrain development in mouse, human, and stem cells. *Cell* 167(2), 566-580.

Examples

```
sce.h.es <- LaMannoBrainData()

sce.h.em <- LaMannoBrainData("human-embryo")

sce.h.ip <- LaMannoBrainData("human-ips")

sce.m.ad <- LaMannoBrainData("mouse-adult")

sce.m.em <- LaMannoBrainData("mouse-embryo")
```

`LawlorPancreasData` *Obtain the Lawlor pancreas data*

Description

Provides the human pancreas single-cell RNA-seq data from Lawlor et al. (2017).

Usage

```
LawlorPancreasData(legacy = FALSE)
```

Arguments

`legacy`	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Column metadata is provided in the same form as supplied in GSE86469. This contains information such as the cell type labels and patient status.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/lawlor-pancreas`.

Value

A SingleCellExperiment object with a single matrix of read counts.

Author(s)

Aaron Lun

References

Lawlor N et al. (2017). Single-cell transcriptomes identify human islet cell signatures and reveal cell-type-specific expression changes in type 2 diabetes. *Genome Res.* 27(2), 208-222.

Examples

```
sce <- LawlorPancreasData()
```

`LedergorMyelomaData` *Obtain the Ledergor Myeloma data*

Description

Obtain the human multiple myeloma single-cell RNA-seq data from Ledergor et al. (2018).

Usage

```
LedergorMyelomaData(ensembl = FALSE, location = TRUE, legacy = FALSE)
```

Arguments

`ensembl`	Logical scalar indicating whether the output row names should contain Ensembl identifiers.
`location`	Logical scalar indicating whether genomic coordinates should be returned.
`legacy`	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Column metadata was created from the sample metadata file in GSE117156. It contains an 'Experiment_ID' column, from which the tissue and subject of origin were extracted, as well as the condition and treatment status of the subject.

If `ensembl=TRUE`, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

If `location=TRUE`, the coordinates of the Ensembl gene models are stored in the `rowRanges` of the output. Note that this is only performed if `ensembl=TRUE`.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/ledergor-myeloma`.

Value

A SingleCellExperiment object with a single matrix of read counts.

Author(s)

Milan Malfait

References

Ledergor G et al. (2018) Single cell dissection of plasma cell heterogeneity in symptomatic and asymptomatic myeloma. *Nat Med.* 24(12), 1867–1876.

Examples

```
sce <- LedergorMyelomaData()
```

`LengESCData` *Obtain the Leng ESC data*

Description

Obtain the human embryonic stem cell single-cell RNA-seq data from Leng et al. (2015).

Usage

```
LengESCData(ensembl = FALSE, location = TRUE, legacy = FALSE)
```

Arguments

`ensembl`	Logical scalar indicating whether gene symbols should be converted to Ensembl annotation.
`location`	Logical scalar indicating whether genomic coordinates should be returned.
`legacy`	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Column metadata contains the cell line, experiment number and experimentally determined cell cycle phase for each cell.

If `ensembl=TRUE`, the gene symbols in the published annotation are converted to Ensembl. If `location=TRUE`, the coordinates of the Ensembl gene models are stored in the `rowRanges` of the output.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/leng-esc`.

Value

A SingleCellExperiment object with a single matrix of normalized expected read counts.

Author(s)

Aaron Lun

References

Leng F et al. (2015). Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq experiments. *Nat. Methods* 12(10), 947-950.

Examples

```
sce <- LengESCData()
```

`listDatasets` *List all available datasets*

Description

Summary information for all available datasets in the **scRNAseq** package.

Usage

```
listDatasets()
```

Details

A study may contribute multiple datasets if they cannot be reasonably combined (e.g., different species). The reported number of cells refers only to the dataset as it is stored in **scRNAseq**; this may be different to the number of cells used by the authors in their analysis, e.g., due to filtering.

Value

A DataFrame where each row corresponds to a dataset, containing the fields:

- `Reference`, a Markdown-formatted citation to `scripts/ref.bib` in the **scRNAseq** installation directory.
- `Taxonomy`, an identifier for the organism.
- `Part`, the part of the organism being studied.
- `Number`, the total number of cells in the dataset.
- `Call`, the relevant R call required to construct the dataset.

Author(s)

Aaron Lun

Examples

```
listDatasets()
```

`listPaths` *List all paths for a dataset*

Description

List the available paths to subdatasets within a version of a given dataset.

Usage

```
listPaths(
  name,
  version,
  package = "scRNAseq",
  cache = cacheDirectory(),
  overwrite = FALSE,
  include.metadata = FALSE
)
```

Arguments

`name`	String containing the name of the dataset.
`version`	String containing the version of the dataset.
`package`	String containing the name of the package.
`cache, overwrite`	Arguments to pass to `saveVersion` or `saveFile`.
`include.metadata`	Logical scalar indicating whether to report the metadata for each subdataset.

Value

If `include.metadata=FALSE`, a character vector containing the paths to subdatasets within the specified version of the dataset. If `name` does not contain any subdatasets, `NA` is returned.

Otherwise, a DFrame is returned containing the metadata for each subdataset, e.g., the title and description. More details can be found in the Bioconductor metadata schema at https://github.com/ArtifactDB/bioconductor-metadata-index.

Author(s)

Aaron Lun

Examples

```
listPaths("he-organs-2020", "2023-12-21")
listPaths("he-organs-2020", "2023-12-21", include.metadata=TRUE)
listPaths("zeisel-brain-2015", "2023-12-14") # no subdatasets
listPaths("zeisel-brain-2015", "2023-12-14", include.metadata=TRUE)
```

listVersions *List available versions*

Description

List the available and latest versions for a dataset.

Usage

```
listVersions(name)

fetchLatestVersion(name)
```

Arguments

name	String containing the name of the dataset.

Value

For `listVersions`, a character vector containing the names of the available versions of the dataset.

For `fetchLatestVersion`, a string containing the name of the latest version.

Author(s)

Aaron Lun

Examples

```
listVersions("zeisel-brain-2015")
fetchLatestVersion("zeisel-brain-2015")
```

LunSpikeInData *Obtain the Lun spike-in data*

Description

Obtain the spike-in single-cell RNA-seq data from Lun et al. (2017).

Usage

```
LunSpikeInData(
  which = c("416b", "tropho"),
  split.oncogene = FALSE,
  location = TRUE,
  legacy = FALSE
)
```

Arguments

`which`	String specifying whether the 416B or trophoblast data should be obtained.
`split.oncogene`	Logical scalar indicating whether the oncogene should be split to a separate `altExp`.
`location`	Logical scalar indicating whether genomic coordinates should be returned.
`legacy`	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Row data contains a single `"Length"` field describing the total exonic length of each feature.

Column metadata is provided in the same form as supplied in E-MTAB-5522. This contains information such as the cell type, plate of origin, spike-in addition order and oncogene induction.

Two sets of spike-ins were added to each cell in each dataset. These are available as the `"SIRV"` and `"ERCC"` entries in the `altExps`.

If `split.oncogene=TRUE` and `which="416b"`, the CBFB-MYH11-mcherry oncogene is moved to extra `"oncogene"` entry in the `altExps`.

If `location=TRUE`, the coordinates of the Ensembl gene models are stored in the `rowRanges` of the output.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/lun-spikein`.

Value

A SingleCellExperiment object with a single matrix of read counts.

Author(s)

Aaron Lun

References

Lun ATL et al. (2017). Assessing the reliability of spike-in normalization for analyses of single-cell RNA sequencing data. *Genome Res.* 27(11), 1795-1806.

Examples

```
sce <- LunSpikeInData()

sce <- LunSpikeInData("tropho")
```

`MacoskoRetinaData` *Obtain the Macosko retina data*

Description

Obtain the mouse retina single-cell RNA-seq data from Macosko et al. (2016).

Usage

```
MacoskoRetinaData(ensembl = FALSE, location = TRUE, legacy = FALSE)
```

Arguments

`ensembl`	Logical scalar indicating whether the output row names should contain Ensembl identifiers.
`location`	Logical scalar indicating whether genomic coordinates should be returned.
`legacy`	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Column metadata contains the cluster identity as reported in the paper. Note that some cells will have `NA` identities as they are present in the count matrix but not in the metadata file. These are presumably low-quality cells that were discarded prior to clustering.

If `ensembl=TRUE`, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

If `location=TRUE`, the coordinates of the Ensembl gene models are stored in the `rowRanges` of the output. Note that this is only performed if `ensembl=TRUE`.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/macosko-retina`.

Value

A SingleCellExperiment object with a single matrix of UMI counts.

Author(s)

Aaron Lun

References

Macosko E et al. (2016). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 161(5), 1202-1214.

Examples

```
sce <- MacoskoRetinaData()
```

`MairPBMCData` *Obtain the Mair CITE-seq data*

Description

Obtain the Mair PBMC targeted CITE-seq data from Mair et al. (2020).

Usage

```
MairPBMCData(
  mode = c("rna", "adt"),
  ensembl = FALSE,
  location = TRUE,
  legacy = FALSE
)
```

Arguments

`mode`	Character vector specifying whether to return either or both the RNA and ADT counts.
`ensembl`	Logical scalar indicating whether the output row names should contain Ensembl identifiers.
`location`	Logical scalar indicating whether genomic coordinates should be returned.
`legacy`	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Column metadata contains the donor identity and cartridge of origin. Some libraries may also be classified as multiplets or have undeterminate origins after hash tag debarcoding.

If `ensembl=TRUE`, the gene symbols in the RNA data are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

If `location=TRUE`, the coordinates of the Ensembl gene models are stored in the `rowRanges` for the RNA data. Note that this is only performed if `ensembl=TRUE`.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/mair-pbmc`.

Value

A SingleCellExperiment object with a single matrix of UMI counts corresponding to the first `mode`, with an optional alternative Experiment if there is a second `mode`.

Author(s)

Stephany Orjuela, with modifications from Aaron Lun

References

Mair C et al. (2020). A targeted multi-omic analysis approach measures protein expression and low-abundance transcripts on the single-cell level. *Cell Rep.* 31, 107499

Examples

```
sce <- MairPBMCData()
```

MarquesBrainData *Obtain the Marques brain data*

Description

Obtain the mouse brain single-cell RNA-seq data from Marques et al. (2016).

Usage

```
MarquesBrainData(ensembl = FALSE, location = TRUE, legacy = FALSE)
```

Arguments

`ensembl`	Logical scalar indicating whether the output row names should contain Ensembl identifiers.
`location`	Logical scalar indicating whether genomic coordinates should be returned.
`legacy`	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Column metadata is provided in the same form as supplied in GSE75330. This contains information such as the cell type and age/sex of the mouse of origin for each cell.

Note that some genes may be present in multiple rows corresponding to different genomic locations. These additional rows are identified by a `_loc[2-9]` suffix in their row names. Users may wish to consider either removing them or merging them, e.g., with `scater::sumCountsAcrossFeatures`.

If `ensembl=TRUE`, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained. All searching is performed after removing the `_loc[2-9]` suffix.

If `location=TRUE`, the coordinates of the Ensembl gene models are stored in the `rowRanges` of the output. Note that this is only performed if `ensembl=TRUE`.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/marques-brain`.

Value

A SingleCellExperiment object with a single matrix of UMI counts.

Author(s)

Aaron Lun

References

Marques A et al. (2016). Oligodendrocyte heterogeneity in the mouse juvenile and adult central nervous system. *Science* 352(6291), 1326-1329.

Examples

```
sce <- MarquesBrainData()
```

`MessmerESCData` *Obtain the Messmer ESC data*

Description

Obtain the human embryonic stem cell single-cell RNA-seq data from Messmer et al. (2019).

Usage

```
MessmerESCData(location = TRUE, legacy = FALSE)
```

Arguments

`location`	Logical scalar indicating whether genomic coordinates should be returned.
`legacy`	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Row data contains a single `"Length"` field describing the total exonic length of each feature.

Column metadata is provided in the same form as supplied in E-MTAB-6819. This contains information such as the cell phenotype (naive or primed) and the batch of origin. Note that counts for technical replicates have already been summed together.

Count data for ERCC spike-ins are stored in the `"ERCC"` entry of the `altExps`.

If `location=TRUE`, the coordinates of the Ensembl gene models are stored in the `rowRanges` of the output.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/messmer-esc`.

Value

A SingleCellExperiment object with a single matrix of read counts.

Author(s)

Aaron Lun

References

Messmer T et al. (2019). Transcriptional heterogeneity in naive and primed human pluripotent stem cells at single-cell resolution. *Cell Rep* 26(4), 815-824.e4

Examples

```
sce <- MessmerESCData()
```

`MuraroPancreasData` *Obtain the Muraro pancreas data*

Description

Obtain the human pancreas single-cell RNA-seq data from Muraro et al. (2016).

Usage

```
MuraroPancreasData(ensembl = FALSE, location = TRUE, legacy = FALSE)
```

Arguments

`ensembl`	Logical scalar indicating whether the output row names should contain Ensembl identifiers.
`location`	Logical scalar indicating whether genomic coordinates should be returned.
`legacy`	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Row data contains fields for the symbol and chromosomal location of each gene.

Column metadata is derived from the columns of the count matrix provided in GSE85241, with additional cell type labels obtained from the authors (indirectly, via the Hemberg group). Some cells have `NA` labels and were presumably removed prior to downstream analyses.

Count data for ERCC spike-ins are stored in the `"ERCC"` entry of the `altExps`.

If `ensembl=TRUE`, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

If `location=TRUE`, the coordinates of the Ensembl gene models are stored in the `rowRanges` of the output. Note that this is only performed if `ensembl=TRUE`.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/muraro-pancreas`.

Value

A SingleCellExperiment object with a single matrix of UMI counts.

Author(s)

Aaron Lun, using additional metadata obtained by Vladimir Kiselev.

References

Muraro MJ et al. (2016). A single-cell transcriptome atlas of the human pancreas. *Cell Syst.* 3(4), 385-394.

Examples

```
sce <- MuraroPancreasData()
```

`NestorowaHSCData` *Obtain the Nestorowa HSC data*

Description

Obtain the mouse haematopoietic stem cell single-cell RNA-seq data from Nestorowa et al. (2015).

Usage

```
NestorowaHSCData(remove.htseq = TRUE, location = TRUE, legacy = FALSE)
```

Arguments

`remove.htseq`	Logical scalar indicating whether HT-seq alignment statistics should be removed.
`location`	Logical scalar indicating whether genomic coordinates should be returned.
`legacy`	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Rows corresponding to HT-seq's alignment statistics are removed by default. These can be retained by setting `remove.htseq=FALSE`.

Column metadata includes the cell type mapping, as described on the website (see References), and the FACS expression levels of selected markers. Note that these are stored as nested matrices within the `colData`.

Diffusion map components are provided as the `"diffusion"` entry in the `reducedDims`.

Counts for ERCC spike-ins are stored in the `"ERCC"` entry in the `altExps`.

If `location=TRUE`, the coordinates of the Ensembl gene models are stored in the `rowRanges` of the output.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/nestorowa-hsc`.

Value

A SingleCellExperiment object with a single matrix of read counts.

Author(s)

Aaron Lun

References

Nestorowa S et al. (2016). A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation *Blood* 128, e20-e31.

Gene and protein expression in adult haematopoiesis: Data. http://blood.stemcells.cam.ac.uk/single_cell_atlas.html#data.

Examples

```
sce <- NestorowaHSCData()
```

NowakowskiCortexData *Obtain the Nowakowski cortex data*

Description

Obtain the human cortex single-cell RNA-seq dataset from Nowakowski et al. (2017).

Usage

```
NowakowskiCortexData(ensembl = FALSE, location = TRUE, legacy = FALSE)
```

Arguments

`ensembl`	Logical scalar indicating whether the output row names should contain Ensembl identifiers.
`location`	Logical scalar indicating whether genomic coordinates should be returned.
`legacy`	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Column metadata includes the presumed cell type (`WGCNAcluster`), patient and tissue region of origin. A variety of dimensionality reduction results are also provided.

If `ensembl=TRUE`, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

If `location=TRUE`, the coordinates of the Ensembl gene models are stored in the `rowRanges` of the output. This is only performed when `ensembl=TRUE`.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/nowakowski-cortex`.

Value

A SingleCellExperiment object with a single matrix of TPMs. The `reducedDims` contains an assortment of dimensionality reduction results.

Author(s)

Aaron Lun

References

Nowakowski S et al. (2017). Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science* 358, 1318-1323.

Examples

```
sce <- NowakowskiCortexData()
```

PaulHSCData *Obtain the Paul HSC data*

Description

Obtain the mouse haematopoietic stem cell single-cell RNA-seq data from Paul et al. (2015).

Usage

```
PaulHSCData(
  ensembl = FALSE,
  discard.multiple = TRUE,
  location = TRUE,
  legacy = FALSE
)
```

Arguments

ensembl	Logical scalar indicating whether the output row names should contain Ensembl identifiers.
discard.multiple	Logical scalar indicating whether ambiguous rows should be discarded.
location	Logical scalar indicating whether genomic coordinates should be returned.
legacy	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Column metadata includes the plate and the mouse of origin, fluoresence intensities from indexed sorting and the number of cells in each well.

Some of the original rownames are concatenated symbols from multiple genes. We consider these rows to represent ambiguously assigned counts and discard them if `discard.multiple=TRUE`.

If `ensembl=TRUE`, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

If `location=TRUE`, the coordinates of the Ensembl gene models are stored in the `rowRanges` of the output. Note that this is only performed if `ensembl=TRUE`.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/nestorowa-hsc`.

Value

A SingleCellExperiment object with a single matrix of read counts.

Author(s)

Aaron Lun

References

Paul F et al. (2015). Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell* 163, 1663-77.

Examples

```
sce <- PaulHSCData()
```

`polishDataset` *Polish dataset for saving*

Description

Prepare a SummarizedExperiment or SingleCellExperiment to be saved with `saveDataset`. This performs some minor changes to improve storage efficiency.

Usage

```
polishDataset(
  x,
  strip.inner.names = TRUE,
  reformat.assay.by.density = 0.3,
  attempt.integer.conversion = TRUE,
  remove.altexp.coldata = TRUE,
  forbid.nested.altexp = TRUE
)
```

Arguments

`x`	A SummarizedExperiment or one of its subclasses.
`strip.inner.names`	Logical scalar indicating whether to strip redundant names from internal objects, e.g., dimnames of the assays, row names of reduced dimensions, column names of alternative experiments. This saves some space in the on-disk representation.
`reformat.assay.by.density`	Numeric scalar indicating whether to optimize assay formats based on the density of non-zero values. Assays with densities above this number are converted to ordinary dense arrays (if they are not already), while those with lower densities are converted to sparse matrices. This can be disabled by setting it to `NULL`.

attempt.integer.conversion
: Logical scalar indicating whether to convert double-precision assays containing integer values to actually have the integer type. This can improve efficiency of downstream applications by avoiding the need to operate in double precision.

remove.altexp.coldata
: Logical scalar indicating whether column data for alternative experiments should be removed. This defaults to TRUE as the alternative experiment column data is usually redundant with that of the main experiment.

forbid.nested.altexp
: Logical scalar indicating whether nested alternative experiments (i.e., alternative experiments of alternative experiments) should be forbidden. This defaults to TRUE as nested alternative experiments are usually the result of some mistake in altExp preparation.

Value

A modified copy of x.

Author(s)

Aaron Lun

Examples

```
mat <- matrix(rpois(1000, lambda=0.2), 100, 10) * 1.0
rownames(mat) <- sprintf("GENE_%i", seq_len(nrow(mat)))
colnames(mat) <- head(LETTERS, 10)

library(SingleCellExperiment)
sce <- SingleCellExperiment(list(counts=mat))
str(assay(sce, withDimnames=FALSE))

polished <- polishDataset(sce)
str(assay(polished, withDimnames=FALSE))
```

PollenGliaData *Obtain the Pollen radial glia data*

Description

Obtain the human radial glia single-cell RNA-seq dataset from Pollen et al. (2017).

Usage

```
PollenGliaData(ensembl = FALSE, location = TRUE, legacy = FALSE)
```

Arguments

ensembl	Logical scalar indicating whether the output row names should contain Ensembl identifiers.
location	Logical scalar indicating whether genomic coordinates should be returned.
legacy	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Column metadata includes the anatomical source, sample of origin, presumed cell type and assorted alignment statistics.

If ensembl=TRUE, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

If location=TRUE, the coordinates of the Ensembl gene models are stored in the rowRanges of the output. This is only performed when ensembl=TRUE.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for scRNAseq/pollen-glia.

Value

A SingleCellExperiment object with a single matrix of read counts.

Author(s)

Aaron Lun

References

Pollen A et al. (2017). Molecular identity of human outer radial glia during cortical development. *Cell* 163, 55-67.

Examples

```
sce <- PollenGliaData()
```

---

reexports *Objects exported from other packages*

Description

These objects are imported from other packages. Follow the links below to see their documentation.

gypsum defineTextQuery, gsc

`ReprocessedAllenData` *Reprocessed single-cell data sets*

Description

Obtain the legacy count matrices for three publicly available single-cell RNA-seq datasets. Raw sequencing data were downloaded from NCBI's SRA or from EBI's ArrayExpress, aligned to the relevant genome build and used to quantify gene expression.

Usage

```
ReprocessedAllenData(
  assays = NULL,
  ensembl = FALSE,
  location = TRUE,
  legacy = FALSE
)

ReprocessedTh2Data(
  assays = NULL,
  ensembl = FALSE,
  location = TRUE,
  legacy = FALSE
)

ReprocessedFluidigmData(
  assays = NULL,
  ensembl = FALSE,
  location = TRUE,
  legacy = FALSE
)
```

Arguments

`assays`	Character vector specifying one or more assays to return. Choices are `"tophat_counts"`, `"cufflinks_fpkm"`, `"rsem_counts"` and `"rsem_tpm"`. If `NULL`, all assays are returned.
`ensembl`	Logical scalar indicating whether the output row names should contain Ensembl identifiers.
`location`	Logical scalar indicating whether genomic coordinates should be returned.
`legacy`	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

ReprocessedFluidigmData returns a dataset of 65 human neural cells from Pollen et al. (2014), each sequenced at high and low coverage (SRA accession SRP041736).

ReprocessedTh2Data returns a dataset of 96 mouse T helper cells from Mahata et al. (2014), obtained from ArrayExpress accession E-MTAB-2512. Spike-in counts are stored in the "ERCC" entry of the altExps.

ReprocessedAllenData return a dataset of 379 mouse brain cells from Tasic et al. (2016). This is a re-processed subset of the data from TasicBrainData, and contains spike-in information stored as in the altExps.

In each dataset, the first columns of the colData are sample quality metrics from FastQC and Picard. The remaining fields were obtained from the original study in their GEO/SRA submission and/or as Supplementary files in the associated publication. These two categories of colData are distinguished by a which_qc element in the metadata, which contains the names of the quality-related columns in each object.

If ensembl=TRUE, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

If location=TRUE, the coordinates of the Ensembl gene models are stored in the rowRanges of the output. Note that this is only performed if ensembl=TRUE.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for scRNAseq/legacy-allen, scRNAseq/legacy-fluidigm or scRNAseq/legacy-th2.

Value

A SingleCellExperiment object containing one or more expression matrices of counts and/or TPMs, depending on assays.

Pre-processing details

FASTQ files were either obtained directly from ArrayExpress, or converted from SRA files (downloaded from the Sequence Read Archive) using the SRA Toolkit.

Reads were aligned with TopHat (v. 2.0.11) to the appropriate reference genome (GRCh38 for human samples, GRCm38 for mouse). RefSeq mouse gene annotation (GCF_000001635.23_GRCm38.p3) was downloaded from NCBI on Dec. 28, 2014. RefSeq human gene annotation (GCF_000001405.28) was downloaded from NCBI on Jun. 22, 2015.

featureCounts (v. 1.4.6-p3) was used to compute gene-level read counts. Cufflinks (v. 2.2.0) was used to compute gene-leve FPKMs. Reads were also mapped to the transcriptome using RSEM (v. 1.2.19) to compute read counts and TPM's.

FastQC (v. 0.10.1) and Picard (v. 1.128) were used to compute sample quality control (QC) metrics. However, no filtering on the QC metrics has been performed for any dataset.

References

Pollen AA et al. (2014). Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* 32(10), 1053-8.

Mahata B et al. (2014). Single-cell RNA sequencing reveals T helper cells synthesizing steroids de novo to contribute to immune homeostasis. *Cell Rep*, 7(4), 1130-42.

Tasic A et al. (2016). Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat. Neurosci.* 19(2), 335-46.

Examples

```
sce <- ReprocessedAllenData()
```

`RichardTCellData` *Obtain the Richard T cell data*

Description

Obtain the mouse CD8+ T cell single-cell RNA-seq data from Richard et al. (2018).

Usage

```
RichardTCellData(location = TRUE, legacy = FALSE)
```

Arguments

`location`	Logical scalar indicating whether genomic coordinates should be returned.
`legacy`	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Column metadata is provided in the same form as supplied in E-MTAB-6051. This contains information such as the stimulus, time after stimulation, age of the mice and sequencing batch.

Count data for ERCC spike-ins are stored in the `"ERCC"` entry of the `altExps`.

If `location=TRUE`, the coordinates of the Ensembl gene models are stored in the `rowRanges` of the output.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/richard-tcell`.

Value

A SingleCellExperiment object with a single matrix of read counts.

Author(s)

Aaron Lun

References

Richard AC et al. (2018). T cell cytolytic capacity is independent of initial stimulation strength. *Nat. Immunol.* 19(8), 849-858.

Examples

```
sce <- RichardTCellData()
```

`RomanovBrainData` *Obtain the Romanov brain data*

Description

Obtain the mouse brain single-cell RNA-seq dataset from Romanov et al. (2017).

Usage

```
RomanovBrainData(ensembl = FALSE, location = TRUE, legacy = FALSE)
```

Arguments

`ensembl`	Logical scalar indicating whether the output row names should contain Ensembl identifiers.
`location`	Logical scalar indicating whether genomic coordinates should be returned.
`legacy`	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Column metadata is provided in the same form as supplied in GSE74672. This contains information such as the reporter gene expressed in each cell, the mouse line, dissection type and so on.

Counts for ERCC spike-ins are stored in the `"ERCC"` entry of the `altExps`. Note that some of the spike-in rows have `NA` observations for some (but not all) cells.

If `ensembl=TRUE`, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

If `location=TRUE`, the coordinates of the Ensembl gene models are stored in the `rowRanges` of the output. Note that this is only performed if `ensembl=TRUE`.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/romanov-brain`.

Value

A SingleCellExperiment object with a single matrix of UMI counts.

Author(s)

Aaron Lun, based on code by Vladimir Kiselev and Tallulah Andrews.

References

Romanov RA et al. (2017). Molecular interrogation of hypothalamic organization reveals distinct dopamine neuronal subtypes. *Nat. Neurosci.* 20, 176-188.

Examples

```
sce <- RomanovBrainData()
```

---

saveDataset *Save a dataset to disk*

Description

Save a single-cell dataset to disk, usually in preparation for upload.

Usage

```
saveDataset(x, path, metadata)
```

Arguments

x	A SummarizedExperiment or one of its subclasses.
path	String containing the path to a new directory in which to save x. Any existing directory is removed before saving x.
metadata	Named list containing metadata for this dataset, see the schema returned by fetchMetadataSchema(). Note that the applications.takane property will be automatically added by this function and does not have to be supplied.

Value

x and its metadata are saved into path, and NULL is invisibly returned.

Author(s)

Aaron Lun

See Also

https://github.com/ArtifactDB/bioconductor-metadata-index, on the expected schema for the metadata.

polishDataset, to polish x before saving it.

uploadDirectory, to upload the saved contents.

Examples

```
library(SingleCellExperiment)
sce <- SingleCellExperiment(list(counts=matrix(rpois(1000, lambda=1), 100, 10)))
rownames(sce) <- sprintf("GENE_%i", seq_len(nrow(sce)))
colnames(sce) <- head(LETTERS, 10)

meta <- list(
    title="My dataset",
    description="This is my dataset",
    taxonomy_id="10090",
    genome="GRCh38",
    sources=list(list(provider="GEO", id="GSE12345")),
    maintainer_name="Shizuka Mogami",
    maintainer_email="mogami.shizuka@765pro.com"
)

tmp <- tempfile()
saveDataset(sce, tmp, meta)
list.files(tmp, recursive=TRUE)
alabaster.base::readObject(tmp)
```

---

searchDatasets *Search dataset metadata*

Description

Search for datasets of interest based on matching text in the associated metadata.

Usage

```
searchDatasets(
  query,
  cache = cacheDirectory(),
  overwrite = FALSE,
  latest = TRUE
)
```

Arguments

query	String containing a query in a human-readable syntax or a gypsum.search.clause, see Examples.
cache, overwrite	Arguments to pass to fetchMetadataDatabase.
latest	Whether to only consider the latest version of each dataset.

Details

The returned DataFrame contains the usual suspects like the title and description for each dataset, the number of rows and columns, the organisms and genome builds involved, whether the dataset has any pre-computed reduced dimensions, and so on. More details can be found in the Bioconductor metadata schema at https://github.com/ArtifactDB/bioconductor-metadata-index.

If a dataset contains multiple subdatasets, each subdataset is reported as a separate row in the DataFrame. The location of subdataset is provided in the `path` column. If a dataset does not contain any subdatasets, the `path` entry will be set to `NA`.

Value

A DataFrame where each row corresponds to a (sub)dataset, containing various columns of metadata. Some columns may be lists to capture 1:many mappings.

Author(s)

Aaron Lun

See Also

`surveyDatasets`, to easily obtain a listing of all available datasets.

`translateTextQuery`, for details on the human-readable query syntax.

Examples

```
searchDatasets("brain")[,c("name", "title")]
searchDatasets("Neuro%")[,c("name", "title")]
searchDatasets("taxonomy_id:10090")[,c("name", "title")]
searchDatasets("(genome: GRCm38 AND neuro%) OR pancrea%")[,c("name", "title")]

# We can also use gypsum search clauses via the gsc() function.
# Here, 'asset' is analogous to the 'name' of the dataset.
searchDatasets(gsc(asset="he-organs-2020"))[,c("path")]
searchDatasets(
    (gsc(asset="%brain%", partial=TRUE) |
    gsc(asset="%neur%", partial=TRUE) |
    gsc(text="%neur%", partial=TRUE) |
    gsc(text="%brain%", partial=TRUE)) &
    gsc(field="taxonomy_id", text="10090")
)[,c("name", "title")]
```

`SegerstolpePancreasData` *Obtain the Segerstolpe pancreas data*

Description

Download the human pancreas single-cell RNA-seq (scRNA-seq) dataset from Segerstolpe et al. (2016)

Usage

```
SegerstolpePancreasData(ensembl = FALSE, location = TRUE, legacy = FALSE)
```

Arguments

`ensembl`	Logical scalar indicating whether the output row names should contain Ensembl identifiers.
`location`	Logical scalar indicating whether genomic coordinates should be returned.
`legacy`	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Row data contains fields for the gene symbol and RefSeq transcript IDs corresponding to each gene. The rows of the output object are named with the symbol, but note that these are not unique.

Column metadata were extracted from the `Characteristics` fields of the SDRF file for ArrayExpress E-MTAB-5061. This contains information such as the cell type labels and patient status.

Count data for ERCC spike-ins are stored in the `"ERCC"` entry of the `altExps`. Estimated numbers of spike-in molecules are provided in the `rowData` of this entry. Note that these concentrations are incorrect for donor H1, as 100 uL of spike-in mixture were added for this donor, rather than 25 uL for all others.

If `ensembl=TRUE`, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

If `location=TRUE`, the coordinates of the Ensembl gene models are stored in the `rowRanges` of the output. Note that this is only performed if `ensembl=TRUE`.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/segerstolpe-pancreas`.

Value

A SingleCellExperiment object with a single matrix of read counts.

Author(s)

Aaron Lun

References

Segerstolpe A et al. (2016). Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metab.* 24(4), 593-607.

Examples

```
sce <- SegerstolpePancreasData()
```

`ShekharRetinaData` *Obtain the Shekhar retina data*

Description

Obtain the mouse retina single-cell RNA-seq dataset from Shekhar et al. (2016).

Usage

```
ShekharRetinaData(ensembl = FALSE, location = TRUE, legacy = FALSE)
```

Arguments

`ensembl`	Logical scalar indicating whether the output row names should contain Ensembl identifiers.
`location`	Logical scalar indicating whether genomic coordinates should be returned.
`legacy`	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Column metadata contains the cluster identities as reported in the paper. Note that some cells will have `NA` identities as they are present in the count matrix but not in the metadata file. These are presumably low-quality cells that were discarded prior to clustering.

If `ensembl=TRUE`, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

If `location=TRUE`, the coordinates of the Ensembl gene models are stored in the `rowRanges` of the output. Note that this is only performed if `ensembl=TRUE`.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/shekhar-retina`.

Value

A SingleCellExperiment object with a single matrix of UMI counts.

Author(s)

Aaron Lun

References

Shekhar K et al. (2016). Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell* 166(5), 1308-1323.

Examples

```
sce <- ShekharRetinaData()
```

StoeckiusHashingData *Obtain the Stoeckius cell hashing data*

Description

Obtain the (mostly human) cell hashing single-cell RNA-seq data from Stoeckius et al. (2018).

Usage

```
StoeckiusHashingData(
  type = c("pbmc", "mixed"),
  mode = NULL,
  ensembl = FALSE,
  location = TRUE,
  strip.metrics = TRUE,
  legacy = FALSE
)
```

Arguments

type	String specifying the dataset to obtain.
mode	String specifying the data modalities to obtain, see Details.
ensembl	Logical scalar indicating whether the output row names should contain Ensembl identifiers.
location	Logical scalar indicating whether genomic coordinates should be returned.
strip.metrics	Logical scalar indicating whether quality control metrics should be removed from the HTO/ADT counts.
legacy	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

When `type="pbmc"`, the `mode` can be one or more of:

- `"human"`, the RNA counts for human genes.
- `"mouse"`, the RNA counts for mouse genes. Present as the PBMC dataset is actually a mixture of human PBMCs and unlabelled mouse cells.
- `"hto"`, the HTO counts.
- `"adt1"`, counts for the first set of ADTs (immunoglobulin controls).
- `"adt2"`, counts for the second set of ADTs (cell type-specific markers).

If `mode=NULL`, the default is to use `"human"`, `"mouse"` and `"hto"`.

When `type="mixed"`, the `mode` can be one or more of:

- `"rna"`, the RNA counts for the genes;
- `"hto"`, the HTO counts.

If `mode=NULL`, the default is to use `"rna"` and `"hto"`.

If `ensembl=TRUE`, gene symbols for the RNA counts are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

If `location=TRUE`, the coordinates of the Ensembl gene models are stored in the `rowRanges` of the output. Note that this is only performed if `ensembl=TRUE` and only for the RNA counts.

For the HTO and ADT matrices, some rows correspond to quality control metrics. If `strip.metrics=TRUE`, these rows are removed so that only data for actual HTOs or ADTs are present.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/nestorowa-hsc`.

Value

A SingleCellExperiment object with a matrix of UMI counts corresponding to the first `mode`, plus any number of alternative Experiments containing the remaining `modes`. If multiple `modes` are specified, the output object only contains the intersection of their column names.

Author(s)

Aaron Lun

References

Stoeckius et al. (2018). Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol.* 19, 224.

Examples

```
sce.pbmc <- StoeckiusHashingData()
sce.pbmc

sce.mixed <- StoeckiusHashingData(type="mixed")
sce.mixed
```

`surveyDatasets` *Survey of dataset metadata*

Description

Metadata survey for all available datasets in the **scRNAseq** package.

Usage

```
surveyDatasets(cache = cacheDirectory(), overwrite = FALSE, latest = TRUE)
```

Arguments

`cache`, `overwrite`	Arguments to pass to `fetchMetadataDatabase`.
`latest`	Whether to only consider the latest version of each dataset.

Details

The returned DataFrame contains the usual suspects like the title and description for each dataset, the number of rows and columns, the organisms and genome builds involved, whether the dataset has any pre-computed reduced dimensions, and so on. More details can be found in the Bioconductor metadata schema at https://github.com/ArtifactDB/bioconductor-metadata-index.

If a dataset contains multiple subdatasets, each subdataset is reported as a separate row in the DataFrame. The location of subdataset is provided in the `path` column. If a dataset does not contain any subdatasets, the `path` entry will be set to `NA`.

Value

A DataFrame where each row corresponds to a (sub)dataset, containing various columns of metadata. Some columns may be lists to capture 1:many mappings.

Author(s)

Aaron Lun

See Also

`searchDatasets`, to search on the metadata for specific datasets.

Examples

```
surveyDatasets()
```

`TasicBrainData` *Obtain the Tasic brain data*

Description

Obtain the mouse brain single-cell RNA-seq data from Tasic et al. (2015).

Usage

```
TasicBrainData(ensembl = FALSE, location = TRUE, legacy = FALSE)
```

Arguments

`ensembl`	Logical scalar indicating whether the output row names should contain Ensembl identifiers.
`location`	Logical scalar indicating whether genomic coordinates should be returned.
`legacy`	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Column metadata is provided in the same form as supplied in GSE71585. This contains information such as the reporter gene expressed in each cell, the mouse line, dissection type and so on.

Count data for ERCC spike-ins are stored in the `"ERCC"` entry of the `altExps`. Note that some of the spike-in rows have `NA` observations for some (but not all) cells.

The last 9 columns (containing `_CTX_` in their names) correspond to no-cell control libraries.

If `ensembl=TRUE`, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

If `location=TRUE`, the coordinates of the Ensembl gene models are stored in the `rowRanges` of the output. Note that this is only performed if `ensembl=TRUE`.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/tasic-brain`.

Value

A SingleCellExperiment object with a single matrix of read counts.

Author(s)

Aaron Lun

References

Tasic A et al. (2016). Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat. Neurosci.* 19(2), 335-46.

Examples

```
sce <- TasicBrainData()
```

`UsoskinBrainData` *Obtain the Usoskin brain data*

Description

Obtain the mouse brain single-cell RNA-seq dataset from Usoskin et al. (2015).

Usage

```
UsoskinBrainData(ensembl = FALSE, location = TRUE, legacy = FALSE)
```

Arguments

`ensembl`	Logical scalar indicating whether the output row names should contain Ensembl identifiers.
`location`	Logical scalar indicating whether genomic coordinates should be returned.
`legacy`	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Column metadata is provided in the same form as supplied in External Table 2 of http://linnarssonlab.org/drg/. This contains information such as the library of origin and the cell type.

The count matrix contains information for repeats, marked with `r_` prefixes in the row names; as well as mitochondrial transcripts, marked with `mt-` prefixes.

If `ensembl=TRUE`, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

If `location=TRUE`, the coordinates of the Ensembl gene models are stored in the `rowRanges` of the output. Note that this is only performed if `ensembl=TRUE`.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/usoskin-brain`.

Value

A SingleCellExperiment object with a single matrix of RPMs.

Author(s)

Aaron Lun

References

Usoskin A et al. (2015). Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nat. Neurosci.* 18(1), 145-53.

Examples

```
sce <- UsoskinBrainData()
```

`WuKidneyData` *Obtain the Wu kidney data*

Description

Obtain the mouse kidney single-nuclei RNA-seq data from Wu et al. (2019).

Usage

```
WuKidneyData(
  mode = c("healthy", "disease"),
  ensembl = FALSE,
  location = TRUE,
  legacy = FALSE
)
```

Arguments

`mode`	String indicating whether to return data for healthy and/or diseased donors.
`ensembl`	Logical scalar indicating whether the output row names should contain Ensembl identifiers.
`location`	Logical scalar indicating whether genomic coordinates should be returned.
`legacy`	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Column metadata includes the single-cell technology and whether they came from a diseased or healthy individual.

If `mode` specifies both healthy and disease donors, counts are only reported for the intersection of genes that are present for both donors. This is because the original count matrices had differences in their annotation.

If ensembl=TRUE, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

If location=TRUE, the coordinates of the Ensembl gene models are stored in the rowRanges of the output. Note that this is only performed if ensembl=TRUE.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for scRNAseq/wu-kidney.

Value

A SingleCellExperiment object with a single matrix of read counts.

Author(s)

Aaron Lun

References

Wu H et al. (2019). Advantages of single-nucleus over single-cell RNA sequencing of adult kidney: rare cell types and novel cell states revealed in fibrosis. *J. Am. Soc. Nephrol.* 30, 23-32.

Examples

```
sce <- WuKidneyData("disease")
```

XinPancreasData *Obtain the Xin pancreas data*

Description

Obtain the human pancreas single-cell RNA-seq dataset from Xin et al. (2016).

Usage

```
XinPancreasData(ensembl = FALSE, location = TRUE, legacy = FALSE)
```

Arguments

ensembl	Logical scalar indicating whether the output row names should contain Ensembl identifiers.
location	Logical scalar indicating whether genomic coordinates should be returned.
legacy	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Row data contains fields for the Entrez ID and symbol for each gene. Column metadata was obtained from the authors (indirectly, via the Hemberg group) and contains information such as the cell type labels and donor status.

If `ensembl=TRUE`, the Entrez IDs are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

If `location=TRUE`, the coordinates of the Ensembl gene models are stored in the `rowRanges` of the output. Note that this is only performed if `ensembl=TRUE`.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/xin-pancreas`.

Value

A SingleCellExperiment object with a single matrix of RPKMs.

Author(s)

Aaron Lun, using additional metadata obtained by Vladimir Kiselev.

References

Xin A et al. (2016). RNA sequencing of single human islet cells reveals type 2 diabetes genes. *Cell Metab.* 24(4), 608-615.

Examples

```
sce <- XinPancreasData()
```

`ZeiselBrainData` *Obtain the Zeisel brain data*

Description

Obtain the mouse brain single-cell RNA-seq dataset from Zeisel et al. (2015).

Usage

```
ZeiselBrainData(ensembl = FALSE, location = TRUE, legacy = FALSE)
```

Arguments

`ensembl`	Logical scalar indicating whether the output row names should contain Ensembl identifiers.
`location`	Logical scalar indicating whether genomic coordinates should be returned.
`legacy`	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Row data contains a single "featureType" field describing the type of each feature (endogenous genes, mitochondrial genes, spike-in transcripts and repeats). Spike-ins and repeats are stored as separate entries in the altExps.

Column metadata is provided in the same form as supplied in http://linnarssonlab.org/cortex/. This contains information such as the cell diameter and the published cell type annotations.

If ensembl=TRUE, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

If location=TRUE, the coordinates of the Ensembl gene models are stored in the rowRanges of the output.

Spike-in metadata is added using ERCCSpikeInConcentrations, with molecule counts computed using a volume of 9 nL per cell at a dilution of 1:20000.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for scRNAseq/zeisel-brain.

Value

A SingleCellExperiment object with a single matrix of UMI counts.

Author(s)

Aaron Lun

References

Zeisel A et al. (2015). Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* 347(6226), 1138-42.

Examples

```
sce <- ZeiselBrainData()
```

ZeiselNervousData *Obtain the Zeisel nervous system data*

Description

Obtain the mouse nervous system single-cell RNA-seq dataset from Zeisel et al. (2018).

Usage

```
ZeiselNervousData(location = TRUE, legacy = FALSE)
```

Arguments

location	Logical scalar indicating whether genomic coordinates should be returned.
legacy	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Row data contains the gene symbol as well as some relevant per-gene statistics, e.g., the squared coefficient of variance, mean, and whether it was selected for downstream analyses.

Column data contains a wide variety of fields including patient-level information, sample-level sequencing statistics and many flavors of cell type classification. Note that many numeric columns may have NA values if they could not be successfully parsed form the source file.

If location=TRUE, the coordinates of the Ensembl gene models are stored in the rowRanges of the output.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for scRNAseq/zeisel-nervous.

Value

A SingleCellExperiment object with a single matrix of UMI counts.

Author(s)

Aaron Lun

References

Zeisel A et al. (2018). Molecular architecture of the mouse nervous system. *Cell* 174(4), 999-1014.

Examples

```
if (.Machine$sizeof.pointer > 4) { # too large for 32-bit machines!
    sce <- ZeiselNervousData()
}
```

ZhaoImmuneLiverData *Obtain the Zhao immune liver data*

Description

Obtain the human liver immune single-cell RNA-seq data from Zhao et al. (2020).

Usage

```
ZhaoImmuneLiverData(location = TRUE, filter = FALSE, legacy = FALSE)
```

Arguments

location	Logical scalar indicating whether genomic coordinates should be returned.
filter	Logical scalar indicating if the filtered subset should be returned.
legacy	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Column metadata contains various cell labels as provided by the authors. Some of these labels required assembly on our part:

- The broad label was assigned to each barcode based on whether that barcode was present in each *_identities.tsv.gz in GSE125188's supplementary files.
- For each cell barcode that was present in one of these files, the fine label was generated from the Group annotations inside that file.

We guessed the sample for each cell by assuming that the GEM group numbers match the order of samples in GSE125188. We also assumed that "donor 4" is a typo, given that the paper only mentions 3 donors.

If location=TRUE, the coordinates of the Ensembl gene models are stored in the rowRanges of the output. Note that this is only performed if ensembl=TRUE.

If filter=TRUE, only cells that have been used in the original analysis are returned. Otherwise, the cells used are specified in the retained column of the colData.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for scRNAseq/zhao-immune-liver.

Value

A SingleCellExperiment object with a single matrix of UMI counts.

Author(s)

Aaron Lun

References

Zhao J et al. (2020). Single-cell RNA sequencing reveals the heterogeneity of liver-resident immune cells in human. *Cell Discov* 6, 22.

Examples

```
sce.zhao <- ZhaoImmuneLiverData()
```

ZhongPrefrontalData *Obtain the Zhong prefrontal cortex data*

Description

Obtain the human prefrontal cortex single-cell RNA-seq dataset from Zhong et al. (2018).

Usage

```
ZhongPrefrontalData(ensembl = FALSE, location = TRUE, legacy = FALSE)
```

Arguments

ensembl	Logical scalar indicating whether the output row names should contain Ensembl identifiers.
location	Logical scalar indicating whether genomic coordinates should be returned.
legacy	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Column metadata is scraped from GEO and includes week of gestation, gender and likely cell type.

If ensembl=TRUE, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

If location=TRUE, the coordinates of the Ensembl gene models are stored in the rowRanges of the output. This is only performed when ensembl=TRUE.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for scRNAseq/zhong-prefrontal.

Value

A SingleCellExperiment object with a single matrix of UMI counts.

Author(s)

Aaron Lun

References

Zhong S et al. (2018). A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex. *Nature* 555, 524-528.

Examples

```
sce <- ZhongPrefrontalData()
```

`ZilionisLungData` *Obtain the Zilionis lung cancer data*

Description

Obtain the human/mouse lung cancer single-cell RNA-seq data from Zilionis et al. (2019).

Usage

```
ZilionisLungData(
  which = c("human", "mouse"),
  ensembl = FALSE,
  location = TRUE,
  filter = FALSE,
  legacy = FALSE
)
```

Arguments

`which`	String specifying the species to get data for.
`ensembl`	Logical scalar indicating whether the output row names should contain Ensembl identifiers.
`location`	Logical scalar indicating whether genomic coordinates should be returned.
`filter`	Logical scalar indicating if the filtered subset should be returned.
`legacy`	Logical scalar indicating whether to pull data from ExperimentHub. By default, we use data from the gypsum backend.

Details

Column metadata is provided and contains information on the library, donor ID/animal ID, replicate and tissue.

If `ensembl=TRUE`, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

If `location=TRUE`, the coordinates of the Ensembl gene models are stored in the `rowRanges` of the output. Note that this is only performed if `ensembl=TRUE`.

If `filter=TRUE`, only cells that have been used in the original analysis are returned. The cells used are specified in the `Used` column of the `colData`.

The `reducedDim` contains coordinates of SPRING representations. This may be filled with `NA`s for SPRING coordinates computed on a subset of cells (specified in `colData`).

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/zilionis-lung`.

Value

A SingleCellExperiment object with a single matrix of read counts.

Author(s)

Jens Preussner

References

Zilionis R et al. (2019). Single-cell transcriptomics of human and mouse lung cancers reveals conserved myeloid populations across individuals and species. *Immunity* 50(5), 1317-1334.

Examples

```
sce.human <- ZilionisLungData()

sce.mouse <- ZilionisLungData("mouse")
```

Index